Research Proposal: Securing Cloud-Based RAG Systems Against Prompt Injection Attacks

Arnav Khinvasara
University of California, San Diego
akhinvasara@ucsd.edu

April 19, 2025

Abstract

This research project aims to identify and analyze security vulnerabilities in Retrieval-Augmented Generation, commonly known as RAG, systems deployed within cloud environments, specifically with a focus on prompt injection attacks targeting vector database components. As more and more organizations deploy RAG systems with vector databases in production environments, these systems are prone to being manipulated during the retrieval process, suffering from vector embedding poisoning, and even having sensitive information extracted through carefully crafted inputs. Current approaches to RAG security primarily rely on 1) input sanitization and 2) basic prompt engineering techniques. These methods are often insufficient for RAG systems using vector databases, as they were generically designed for LLM's, and have unique vulnerabilities due to their semantic search mechanisms. This project proposes a three-layer defense framework specifically designed for cloud-based RAG systems with vector databases: pre-retrieval defense, vector database security, and post-retrieval defense. Theoretically, this should allow RAG systems to evade from any of the three main adversarial attempts to manipulate them in cloud environments. The research will address RAG-specific vulnerabilities at the vector database level and develop security approaches that protect the semantic retrieval mechanisms that distinguish RAG systems from traditional LLMs. The project has direct security applications for organizations implementing RAG systems in production environments, which is becoming increasingly common among all industries.

1 Problem Statement

This research project aims to identify and analyze security vulnerabilities in Retrieval-Augmented Generation, commonly known as RAG, systems deployed within cloud environments, specifically with a focus on prompt injection attacks targeting vector database components. As more and more organizations deploy RAG systems with vector databases in production environments, these systems are prone to being manipulated during the retrieval process, suffering from vector embedding poisoning, and even having sensitive information extracted through carefully crafted inputs.

2 Current State and Limitations

Input sanitization and basic prompt engineering techniques that were designed for general Large Language Model (LLM) applications, are the current approaches to RAG security. These methods are often insufficient for RAG systems using vector databases, which have unique vulnerabilities due to their semantic search mechanisms. Additionally, existing research primarily focuses on LLM security in general, with limited attention to the specific security challenges presented by vector database components in cloud environments. Research by Wei et al. [**?**] demonstrated that RAG systems can be vulnerable to attacks that poison the retrieval process, yet defense mechanisms for vector database layers remain underdeveloped and incomprehensive. Although vector databases like Pinecone and Weaviate offer basic security features, the lack of specialized protections against semantic search manipulation and embedding-level attacks, causes a breach within these RAG systems.

3 Novel Approach

This project proposes a three-layer defense framework specifically designed for cloud-based RAG systems with vector databases:

1. Pre-retrieval defense: Enhanced input validation and prompt sanitization tailored to RAG-specific attack vectors.
2. Vector database security: Novel techniques to prevent embedding poisoning, similarity threshold attacks, and metadata filtering bypasses in the vector database layer.
3. Post-retrieval defense: Response verification techniques that detect and mitigate the effects of successful attacks.

The research will be successful because it addresses RAG-specific vulnerabilities at the vector database level rather than applying generic LLM security measures. By focusing on cloud deployment scenarios with vector databases (e.g., Pinecone, Weaviate, or even self-hosted solutions on cloud infrastructure), we can develop security approaches that protect the semantic retrieval mechanisms that distinguish RAG systems from traditional LLMs.

4 Security Applications

The project has direct security applications for organizations implementing RAG systems in production environments. It will provide:

- A threat model specific to cloud-based RAG systems.
- A set of defensive techniques with empirical evaluation.
- Implementation guidelines for securing RAG deployments in cloud environments.
- Metrics for evaluating RAG system security posture.

5 Importance

As RAG systems become fundamental components of enterprise AI strategies, their security implications grow in significance. Vulnerable RAG systems could leak sensitive information, provide harmful responses, or be manipulated to bypass security controls. This research addresses a critical gap in the current understanding of AI system security at the intersection of LLMs and information retrieval systems in cloud environments.

6 Risks and Mitigation Strategies

Scope management will be critical for quality results which is why we'll focus on only a specific subset of prompt injection attacks rather than attempting to address all possible attack vectors. Additionally, implementing cloud-based RAG systems with security controls involves significant technical challenges so we'll use established frameworks (e.g., LangChain, AWS/Azure services) to reduce implementation overhead. We will be building this as a proof of concept rather than a complete application. Latly, measuring defense effectiveness can be subjective so we'll develop quantitative metrics based on attack success rates and response quality degradation.

7 References

References

[1] Chao, J., et al. (2023). "Jailbreaking Black Box Large Language Models in Twenty Queries." *arXiv preprint arXiv:2310.08419.*

[2] Greshake, K., et al. (2023). "More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models." *arXiv preprint arXiv:2302.12173.*

[3] Zou, A., et al. (2023). "Universal and Transferable Adversarial Attacks on Aligned Language Models." *arXiv preprint arXiv:2307.15043.*

[4] Li, A., Zhou, Y., Raghuram, V., Goldstein, T., Goldblum, M. (2025). Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks. *arXiv preprint arXiv:2502.08586.*

[5] Taipalus, T. (2024). Vector database management systems: Fundamental concepts, use-cases, and current challenges. *Cognitive Systems Research*, *85*, 101216.

[6] Zou, W., Geng, R., Wang, B., Jia, J. (2024). PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. *arXiv preprint arXiv:2402.07867.*